Learning molecular models from data

January 14 2014, by Christopher Sciacca

Dr. Heinz Koeppl is part of a new team of scientists at IBM's Zurich research lab focused on systems biology and he is not afraid to claim that one day, soon, advanced biological processes, like cell mitosis, will be represented in mathematical expressions and/or computer code. His new paper in *Nature Methods* explains progress in this space based on his recent work with the tasty fungi known as yeast.

To simplify your paper, is this research taking us closer towards using virtual biological simulations instead of actual experiments?

Indeed. Machine learning techniques as proposed in our paper are essential to get us closer to realistic simulations of cellular molecular processes. Using molecular data, they provide otherwise experimentally inaccessible quantities, such as in vivo binding kinetics of proteins. Having such a kinetic characterization of a process is a prerequisite for realistic simulations that can be used for prediction or hypothesis generation.

Why did you choose yeast for your example?

Yeast is one of the few "model organisms" where a lot of genetic tricks are well established. In particular, we had to engineer yeast to include a synthetic expression system that is well isolated from the host processes and which can serve as a showcase of how well such a kinetic characterization can be done. Even though yeast appears dumb and

simple, it is an eucaryotic cell, which means it includes a nucleus and other structures with many complex signaling pathways, that are actually also found in with human cells.

There are skeptics who believe it is impossible to represent biology as mathematical expressions. What is your counter argument?

I am too much reductionist to be able to take such concerns serious. Why should it not be representable by mathematical expressions or computer code? What is indeed problematic is that our ability to accurately measure cellular processes is - and will be in the near future - quite limited, when compared to the complexity of the already known components of such processes. Not to mention the complexity of the yet to be discovered components. Hence, the inverse problem that our machine learning algorithms have to solve is extremely ill-posed.

So, the crucial question is: when will experimental techniques be advanced enough to allow for a robust reconstruction of cellular processes from data? In contrast to some people, I do not see a fundamental limitation of such an approach. The reconstruction will improve, along with the data quality.

What needs to happen next for this research to reach the next level? Is it all dependent on exascale computing?

I see the major bottleneck not in compute power but in the current number of unknowns in our molecular computational models. Thus, we are limited by experimental techniques and dedicated machine learning algorithms. However, in order to extract the maximal amount of

information in the available data, we focus on exact algorithms such that no artifacts are introduced into models just due to approximations done in the learning algorithms. Such algorithms are often computationally demanding such that even for our modestly complex models, we relied on parallelization. Nevertheless, I currently do not see our research depending on exascale computing.

What is next for your research?

For now, we focused on the kinetic characterization of molecular models in situations where the interaction topology is known. The more challenging problem is to develop machine learning algorithms that can learn topology and kinetic parameters from data.

This field of reverse-engineering molecular networks has received a lot of attention in recent years. However, little work has been done in the reverse engineering of networks from multivariate single-cell data, such as mass cytometry. In the upcoming months we will work on this problem statement.

You are now part of a new emerging computational biology team at IBM Research - Zurich. What are your goals?

The main research thread of this new team will be reverse-engineering algorithms. With onsite expertise in computational biochemistry, mathematical optimization and high performance computing, we are in a good position to advance the reverse-engineering field and finally put it to use for biologists in academia and pharmaceutical companies. Predicting new molecular interactions from experimental data can become the major discovery tool for experimental biology.

More information: "Scalable inference of heterogeneous reaction kinetics from pooled single-cell recordings." Christoph Zechner, Michael Unger, Serge Pelet, Matthias Peter, Heinz Koeppl. *Nature Methods* (2014) DOI: 10.1038/nmeth.2794. Received 05 August 2013 Accepted 08 November 2013 Published online 12 January 2014

Provided by IBM